Natural Language Processing (NLP) in a Nutshell

4. Data Governance Konferenz

Markus Steindl

RISC Software GmbH

JKU JOHANNES KEPLER UNIVERSITÄT LINZ

Member of UAR INNOVATION NETWORK

ober österreich

AnnaLyze

Version 1.1 08/2021

Über die RISC Software GmbH

Grundlagenforschung
in Symbolic Computation
Vorstand: Prof. Peter Paule
Gründer Prof. Bruno Buchberger
60 Mitglieder
1987

1990
Gründung des **Softwareparks Hagenberg** unter Leitung von RISC

Gründung
RISC Software GmbH
(Prof. Bruno Buchberger)
1992

1995
RISCSW spezialisiert sich auf Software für **Logistik und Produktion**

RISC Software GmbH
100%-Tochter der JKU
2004

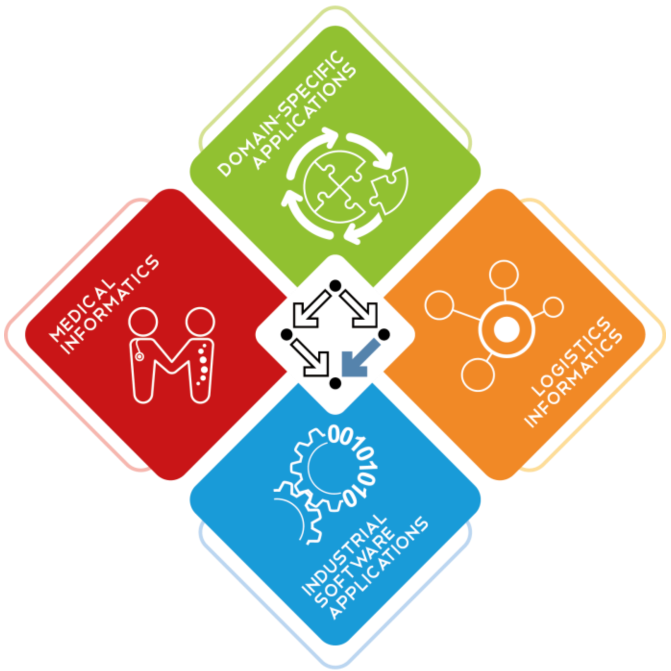
2008
Eingliederung **Medizin-Informatik** und Beteiligung Land OÖ mit 20 %

Gründung der Unit **Domain-specific Applications**
2021

Softwareentwicklung
Angewandte Forschung
(Algorithmische Mathematik)
Technologietransfer

Eigentumsverhältnisse:
80% Johannes Kepler Universität Linz und 20% Upper Austrian Research GmbH (Land OÖ)

RISC Software GmbH - Units



Domain-specific Applications

Agile Softwareentwicklung für Daten- und Prozessmanagement

Logistics Informatics

Software zur Planung, Optimierung, Simulation und Steuerung von Prozessen

Industrial Software Applications

Software für Simulationen, Analysen und Optimierungen in technischen Disziplinen

Medical Informatics

Hochspezialisierte Software für die moderne Medizin

GS1 Workflow-Tool

GS1 Austria

Workflow-Tool zur Abbildung des Prüfprozesses
Import der Artikeldaten via XML-Parser

Eurotrans

DS AUTOMOTION

Leittechnik-Software frei-navigierende fahrerlose Transportsysteme.
Software zur Modellierung der Fahrkurse und Anlagen.
Integrierte Simulationsumgebung für fahrerlose Transportsysteme.

Crashguard

WFL MILLTURN TECHNOLOGIES

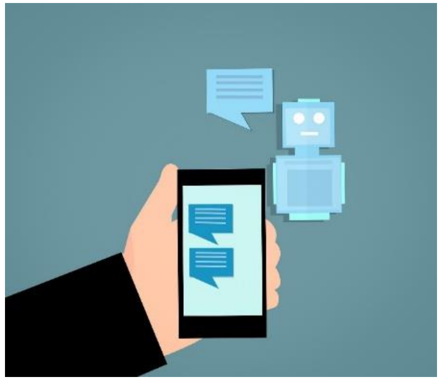
Simulation und Programmierung
von Komplettbearbeitungszentren zur Kollisionsvermeidung

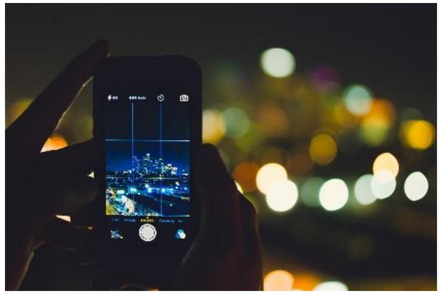
MEDUSA „Leitprojekt Medizintechnik“

Chirurgischer Simulator für zerebrale Aneurysmen zur Ausbildung und OP-Vorbereitung

RISC Software GmbH   JKU JOHANNES KEPLER UNIVERSITY LINZ   Member of UAR INNOVATION NETWORK   upper austria

How does AI affect our lives?





NETFLIX

Spotify

AI

- Recommen-dation Systems
- Asisstance Systems
- Assisted Driving
- Medicine
- Images
- Chatbots
- Smart Home

amazon alexa

Cortana

https://www.inverse.com/article/50031-elon-musk-pulls-self-driving-from-tesla-model-3-options-after-confusion, accessed on 02/12/2019

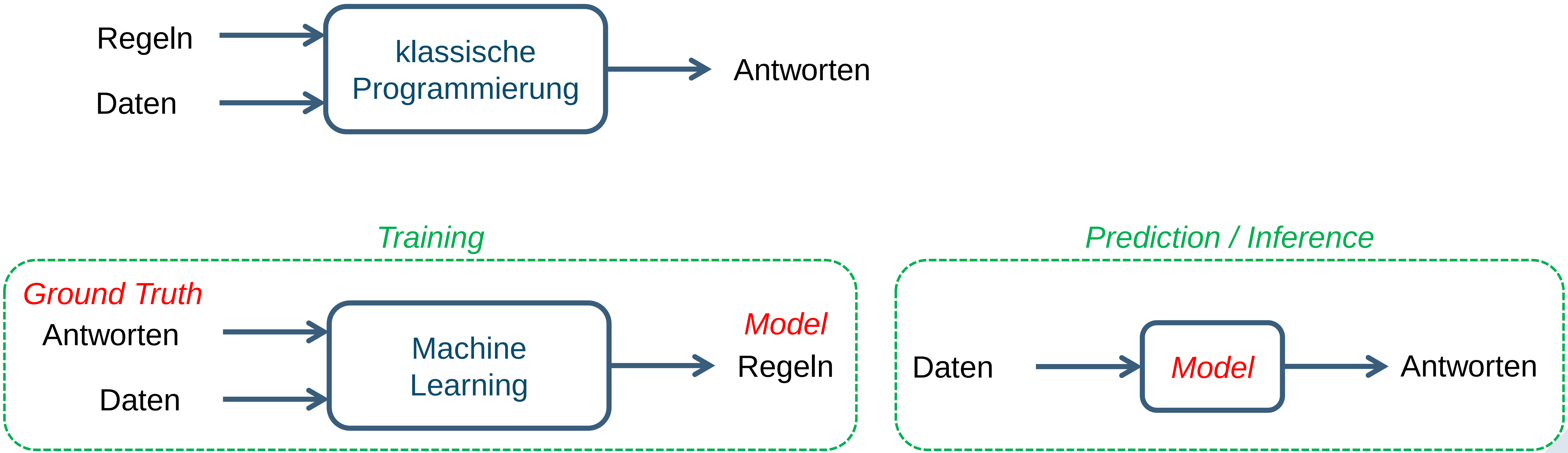
Data & Machine Learning

Anna Lyze

Natural
Language
Processing

RISC Software GmbH

Supervised Machine Learning



Supervised Machine Learning

Image classifier

Data	Answer (Label)
	cat
	dog
	bear

Sentiment classifier

Data	Answer (Label)
“This is the worst product I’ve ever …”	bad
“… quite ok…”	neutral
“… best product for this price …”	good

Supervised Machine Learning

How much labeled training data do we need?

- 1000 to 100,000 images
  for image classification

- 1000 to 100,000 labelled documents
  for text classification

Self-supervised Machine Learning

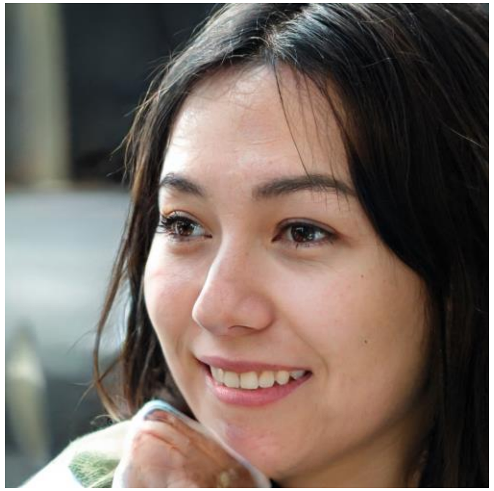
- Labels are created automatically from data

- Ex. 1: Language modeling by *next word prediction*
  - Trains model on large text corpus
  - “$\underbrace{\text{The student closed her}}_{\text{data}}\ \underbrace{\text{book}}_{\text{label}}$”
  - “The apple fell from the _____”
  - Used for training GPT-2, GPT-3, …

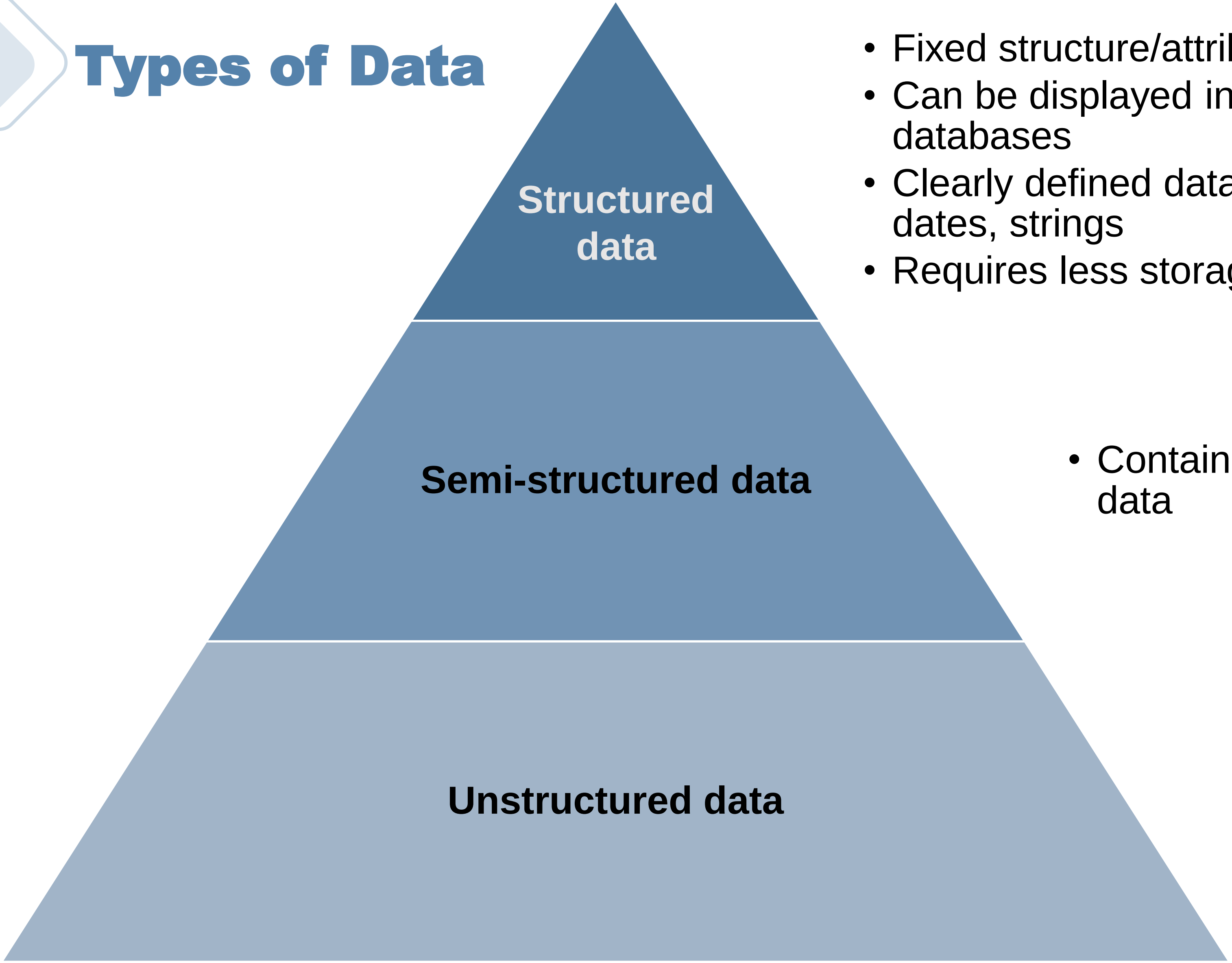
RISC Software GmbH
JKU JOHANNES KEPLER UNIVERSITY LINZ
Member of UAR INNOVATION NETWORK
upper austria

Self-supervised Machine Learning

- Ex. 2: Generative Adversarial Networks (GAN)
  - https://thispersondoesnotexist.com/


Types of Data



- Fixed structure/attributes
- Can be displayed in rows, columns, relational databases
- Clearly defined datatypes such as numbers, dates, strings
- Requires less storage

- Contains structured and unstructured data

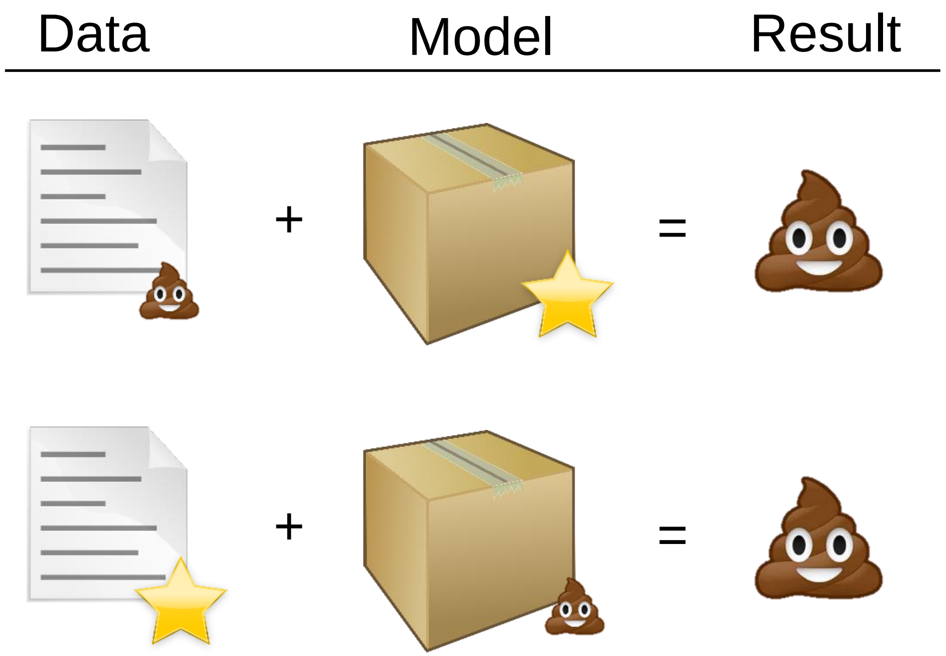
- No formalized structure
- Requires more storage

Why is data preparation so important?

- Incomplete data
- Inconsistent data
- Noise in data

GIGO-paradigm

Garbage In – Garbage Out

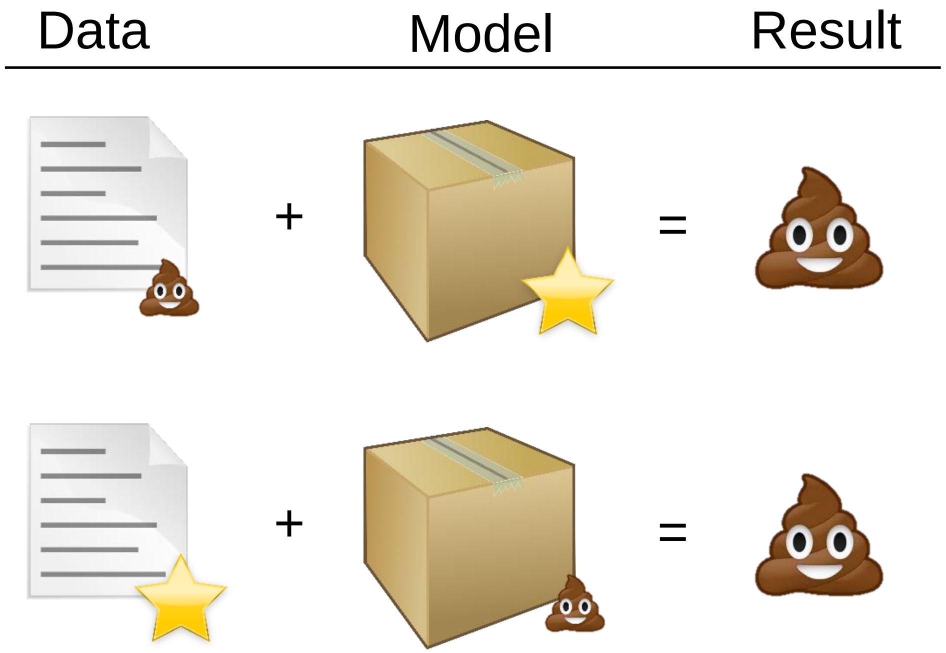


Why is data preparation so important?

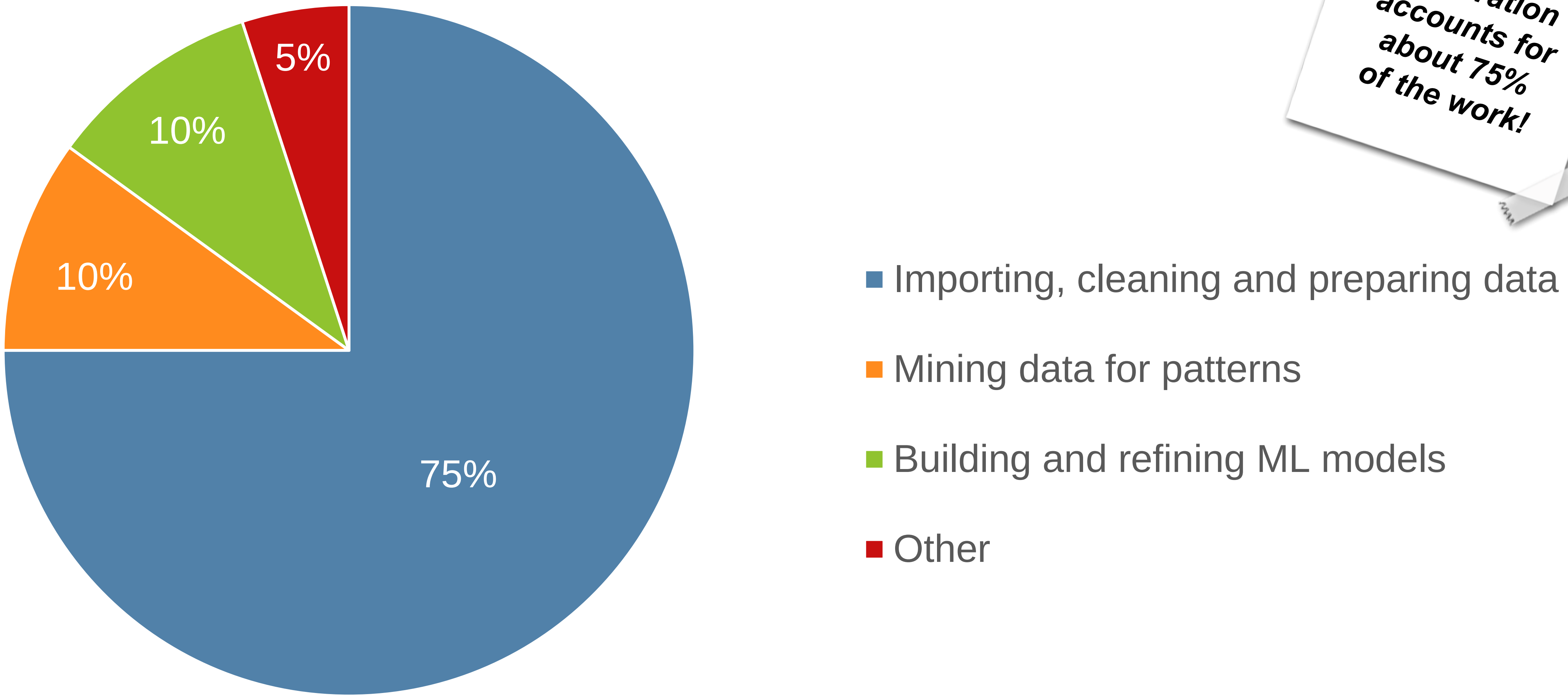
Data quality is crucial for the success of any ML model!

GIGO-paradigm

Garbage In – Garbage Out



Life of a Data Scientist



Data preparation accounts for about 75% of the work!

RISC Software GmbH

Natural Language Processing

AnnaLyze

Natural Language Processing

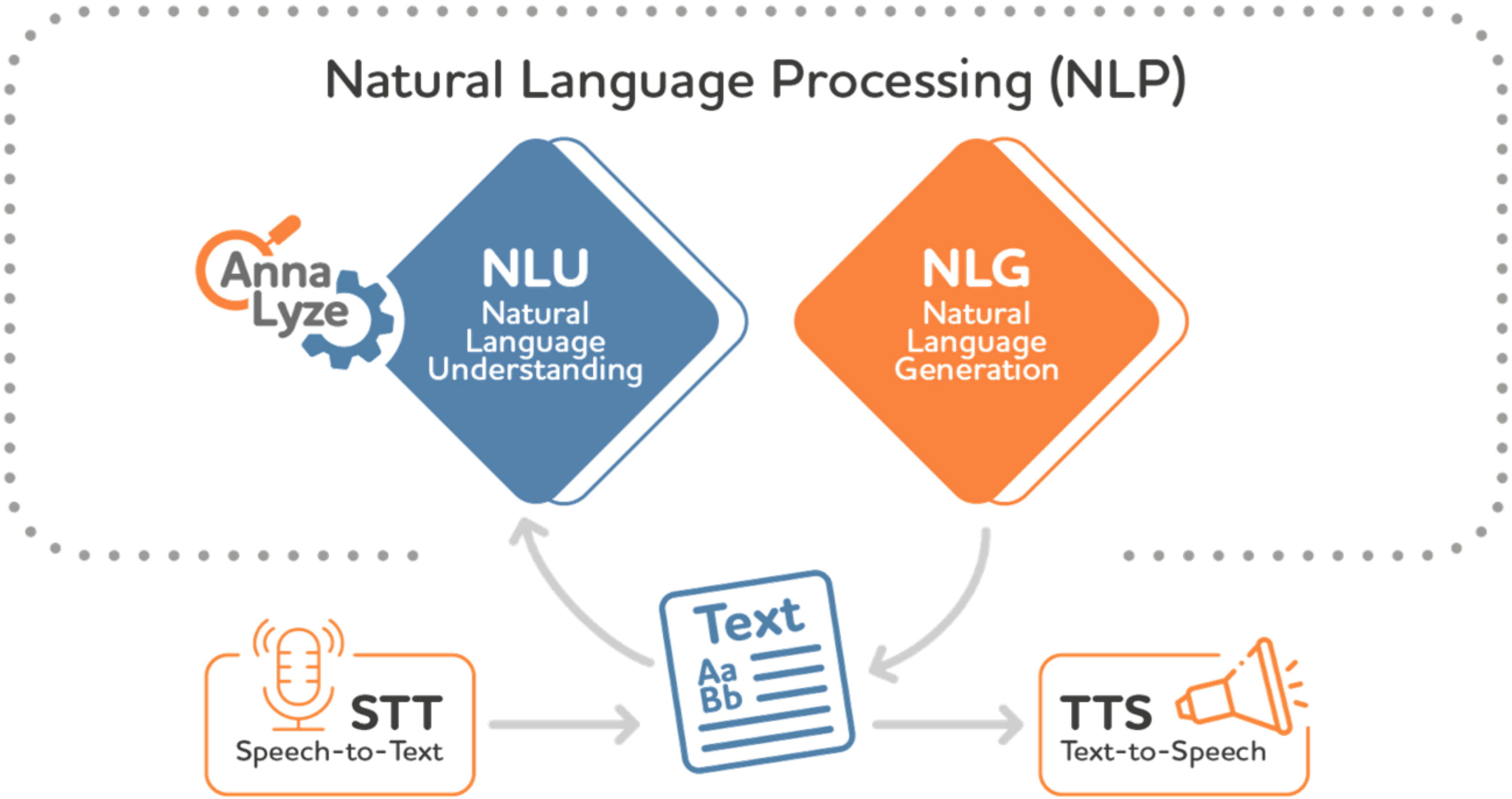
RISC Software GmbH

JKU JOHANNES KEPLER UNIVERSITY LINZ

Member of UAR INNOVATION NETWORK

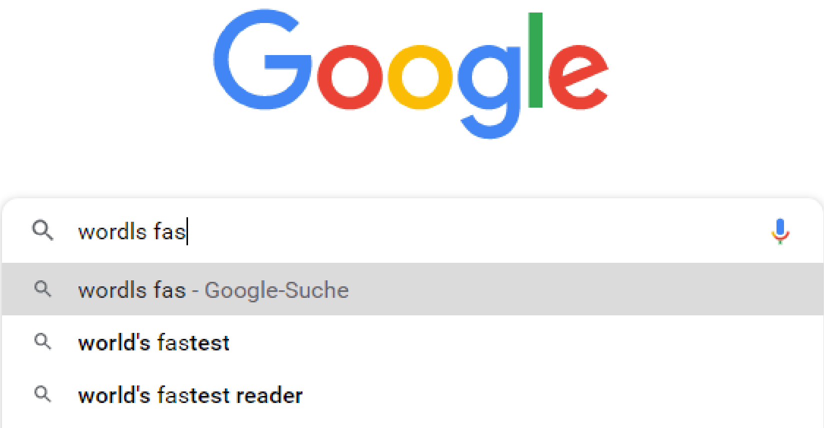
upper austria

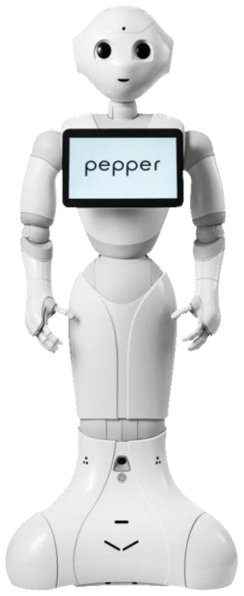
Overview


Natural Language Processing (NLP)
Anna Lyze
NLU
Natural Language Understanding
NLG
Natural Language Generation
Text
Aa Bb
STT
Speech-to-Text
TTS
Text-to-Speech

(Everyday) Applications

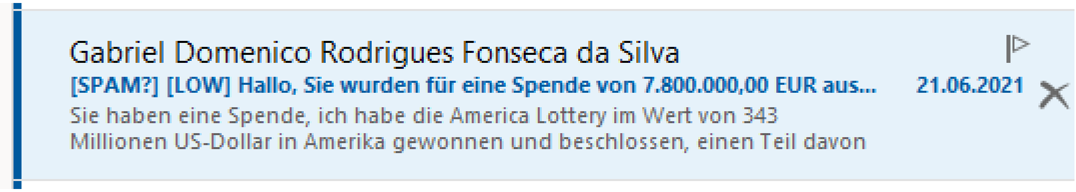
- Search engines
- Jeopardy! IBM Watson
- Customer service robot Pepper
- Email spam filter
- Spell check, autocorrection, autocomplete
- Translation programs
- Voice assistants, e.g. Alexa, Siri und co.
- …





[1]

[2]



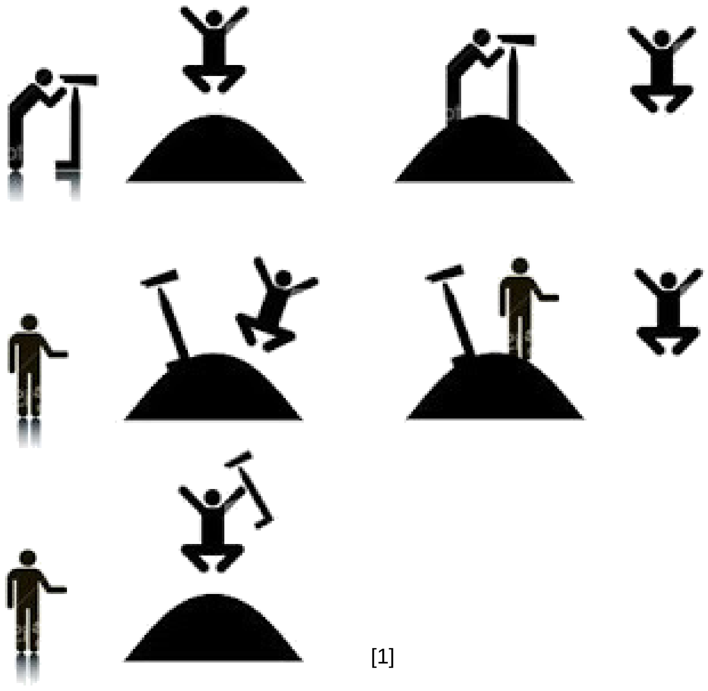
Rechtschreibunk ist gar nicht so wichtik.

Was zählt ist die Mässätsch.

[1] https://www.softbankrobotics.com/emea/de/pepper
[2] https://www.deepl.com/

Why NLP is so challenging

„I saw the man on the hill with a telescope."



[1]

[2]

„Man bit dog"
vs.
„Dog bit man"

[3]

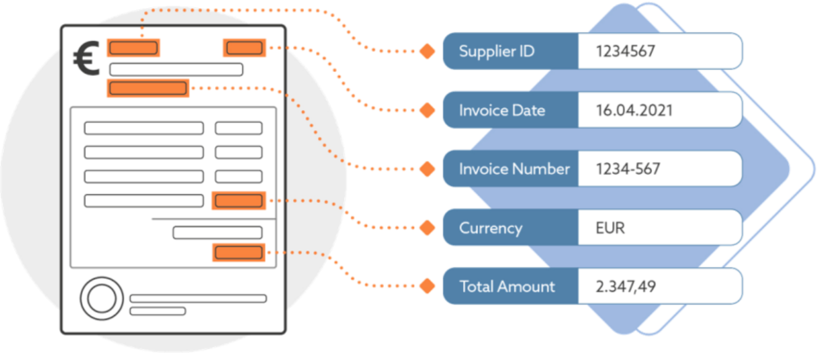
„The spirit is willing, but the flesh is weak."

Russian

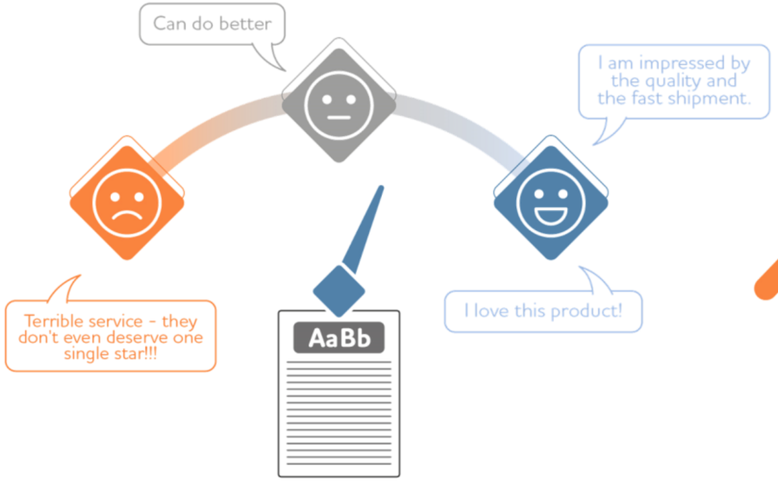
„The vodka is good, but the meat is rotten."

[1] https://allthingslinguistic.com/post/52411342274/how-many-meanings-can-you-get-for-the-sentence-i
[2] https://www.stltoday.com/opinion/columnists/horrigan-man-bites-dog-and-then-the-fun-begins/article_5ee0169c-ac98-5824-bca0-1b944c7d8606.html
[3] https://indianexpress.com/article/india/india-others/dog-bites-man-man-attacks-dog-five-injured-in-scuffle-over-stray/

Common NLP Tasks



information extraction



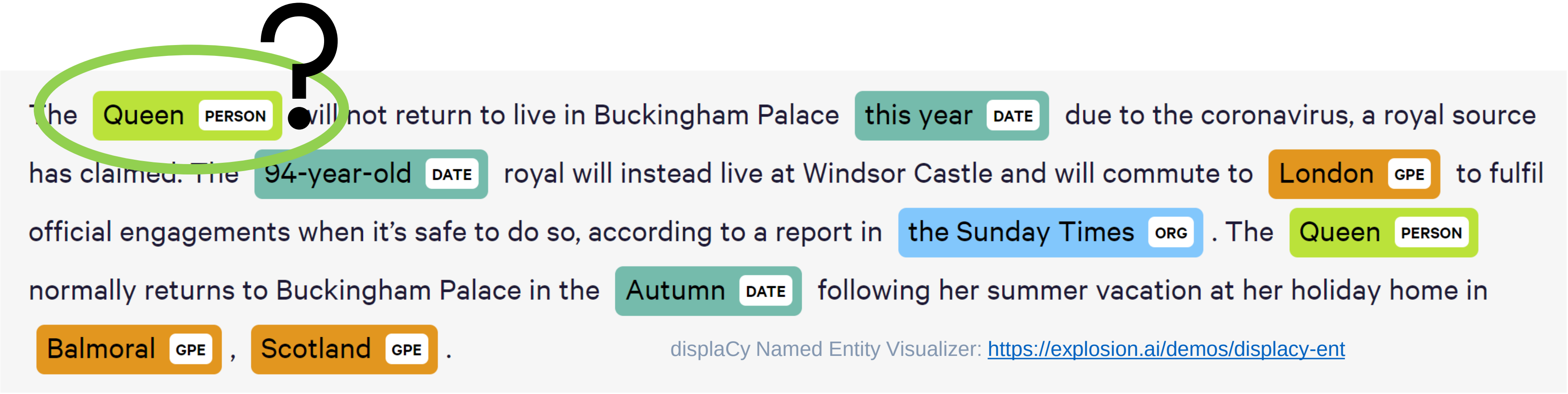
sentiment analysis

text classification

- topic modeling
- automatic summarization
- machine translation
- text generation
- chatbots
- …

Information Extraction – Named Entity recognition (NER)

- Tag named entities
  - persons
  - locations
  - dates
  - etc.



displaCy Named Entity Visualizer: https://explosion.ai/demos/displacy-ent

(Neural) Language Models

- Base models are trained neural nets
  - self-supervised
  - on a large corpus, e.g. 5000 books
  - to do a “simple” task, e.g. next word prediction

- This base model then
  - “understands” language and context
  - can be fine-tuned for NLP-tasks

- Base and fine-tuned models publicly available
  - BERT and T5 by Google
  - GPT-2 by OpenAI
  - Huggingface library

[1]

[1] https://medium.com/@noa.kel/using-bert-with-pytorch-b9624edcda4e

Language Models on Huggingface

- https://huggingface.co/models
- 20,000 models, free-to-use
  - Question Answering
  - Summarization
  - Text Classification
  - Text Generation
  - Text2Text Generation
  - Token Classification
  - Translation
  - Sentence Similarity
  - etc.
- 1800 datasets
  - a good dataset has at least 10,000 or 100,000 labeled texts


[1]

[1] https://medium.com/@noa.kel/using-bert-with-pytorch-b9624edcda4e

NLP-Pipeline: Text Preprocessing

- Normalization
  - Remove special characters
  - Expand abbreviations — isn't → is not
  - Map similar spelling — non-linear → nonlinear
  - Stemming — running → run
  - lemmatization — better → good
  - Stopword removal — the, a, is, are, …
  - Out-of-vocabulary (oov) words
- Tokenization
  - tokens are atomic units
  - Let's go to Leicestershire!

→ | Let | ' | s | go | to | Leicester | ##shire | ! |

There is no unique and always-right preprocessing, as it always depends on the objective and the methods used.

Machines work with numbers, not text.

NLP-Pipeline: Text Encoding

- Basic vectorization approaches
  - **One-Hot-Encoding**
  - Bag of Words (BoW)
  - Bag of N-Grams (BoN)
  - TF-IDF
- Embeddings (e.g. words or documents)

id	word	encoding
1	aaron	[1, 0, 0, ..., 0]
2	ability	[0, 1, 0, ..., 0]
...	...	...
9999	zinc	[0,..., 0, 1, 0]
10000	zone	[0,..., 0, 0, 1]

„The curse of dimensionality“

NLP-Pipeline: Text Encoding

- Basic vectorization approaches
  - One-Hot-Encoding
  - **Bag of Words (BoW)**
  - Bag of N-Grams (BoN)
  - TF-IDF
- Embeddings (e.g. words or documents)

Review 1: „I really love this great film“
Review 2: „this film is really, really great“

↓

„this“, „film“, „is“, „really“, „great“, „I“, „love“

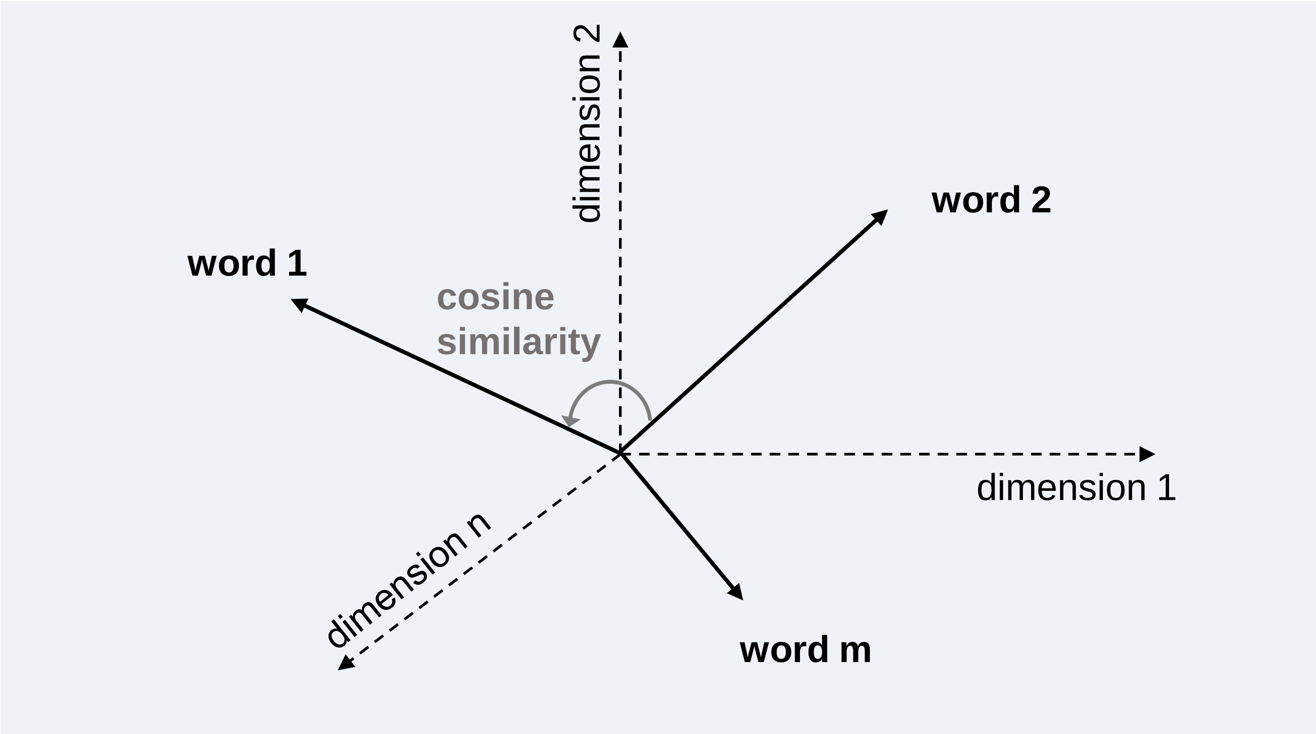
↓

	this	film	is	really	great	I	love
Review 1	1	1	0	1	1	1	1
Review 2	1	1	1	2	1	0	0

„The curse of dimensionality“

NLP-Pipeline: Text Encoding

- Basic vectorization approaches
  - One-Hot-Encoding
  - Bag of Words (BoW)
  - Bag of N-Grams (BoN)
  - TF-IDF
- **Embeddings** (e.g. words or documents)



NLP-Pipeline: Text Encoding

- Basic vectorization approaches
  - One-Hot-Encoding
  - Bag of Words (BoW)
  - Bag of N-Grams (BoN)
  - TF-IDF
- **Embeddings** (e.g. words or documents)

Can you solve it?

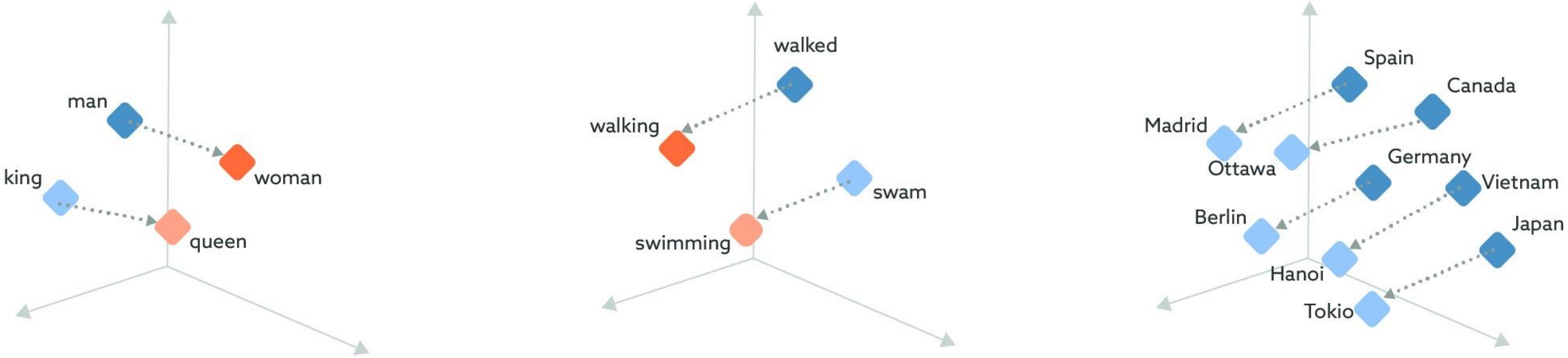
[king] – [man] + [woman] ~= ?

[walking] - [swimming] + [swam] ~= ?

[madrid] - [spain] + [france] ~= ?

Word embeddings

- Famous implementations
  - Word2Vec, GloVe, ...
- Similar embedding ↔ similar meaning



Male — Female   Verb Tense   Country — Capital

[1] https://towardsdatascience.com/legal-applications-of-neural-word-embeddings-556b7515012f

Machines work with numbers, not text.

Embeddings are vectors that encode words and the relations between them. They are state-of-the-art encodings for neural net inputs.

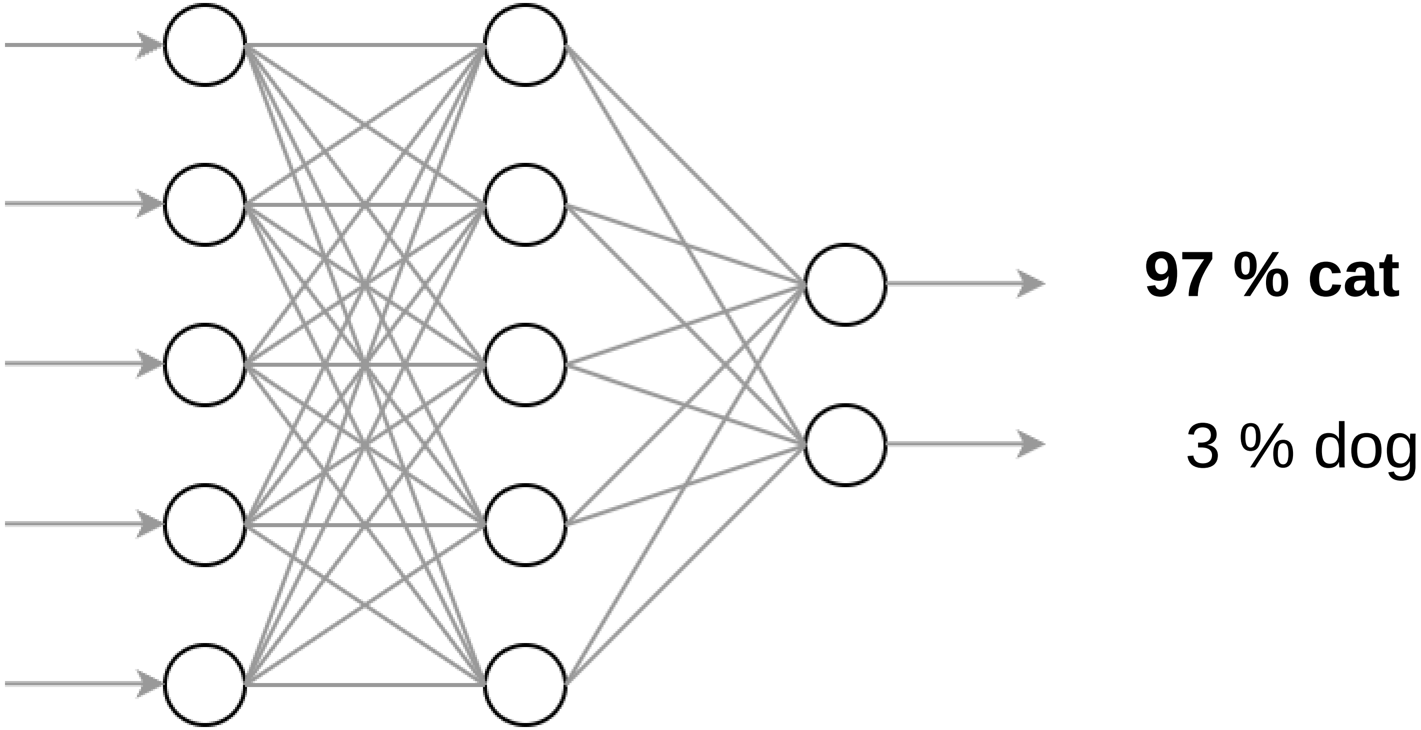
RISC Software GmbH

JOHANNES KEPLER UNIVERSITY LINZ

Member of UAR INNOVATION NETWORK

upper austria

Neural net as classifier



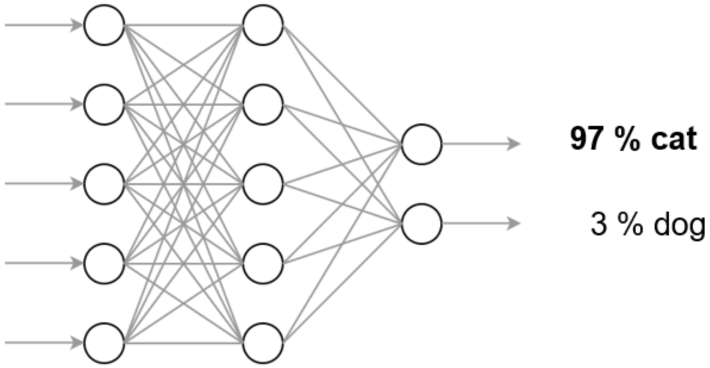
Processing sequences

- Input size is fixed
  for feed forward nets

- E.g. net requires 224 by 224 pixels

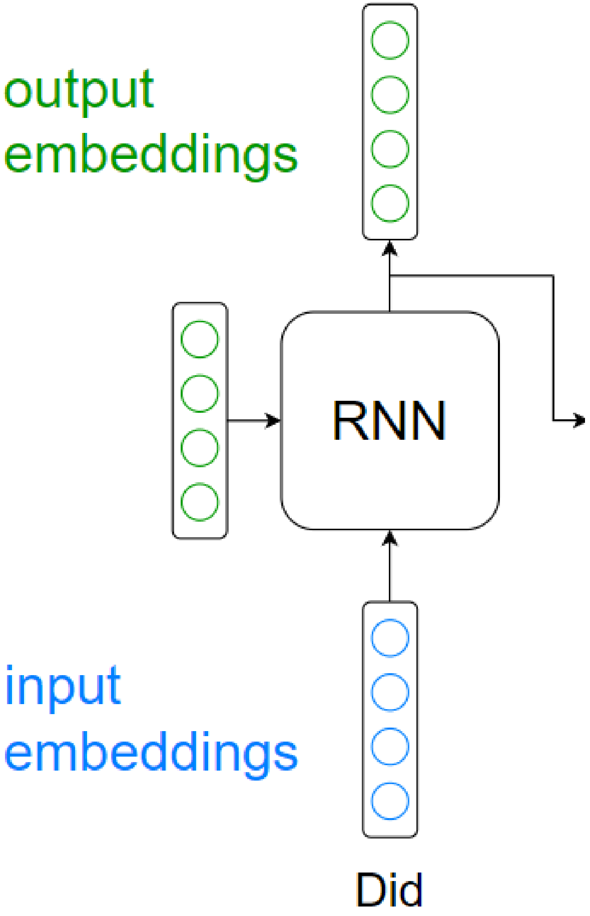
- What if we want to process
  sequences of variable lengths?

- Recurrent neural networks (RNN)
  - Variable input lengths
  - Long short-term memory (LSTM)
  - Gated recurrent unit (GRU)

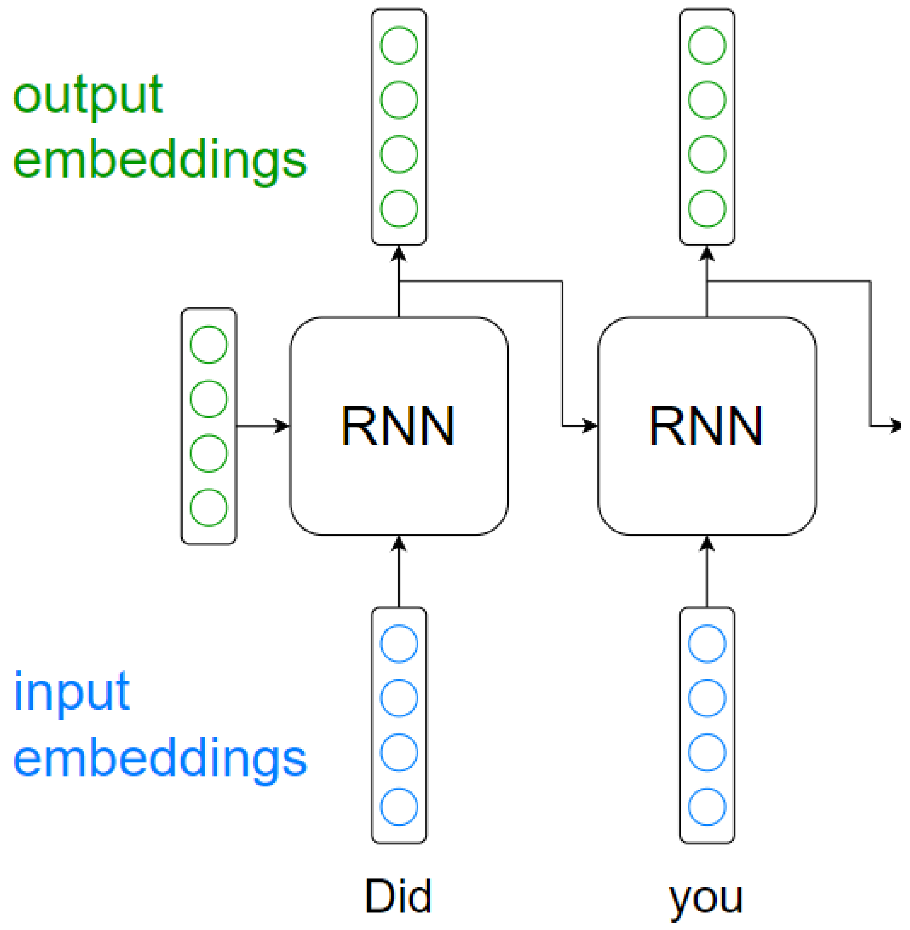


https://en.wikipedia.org/wiki/Cat

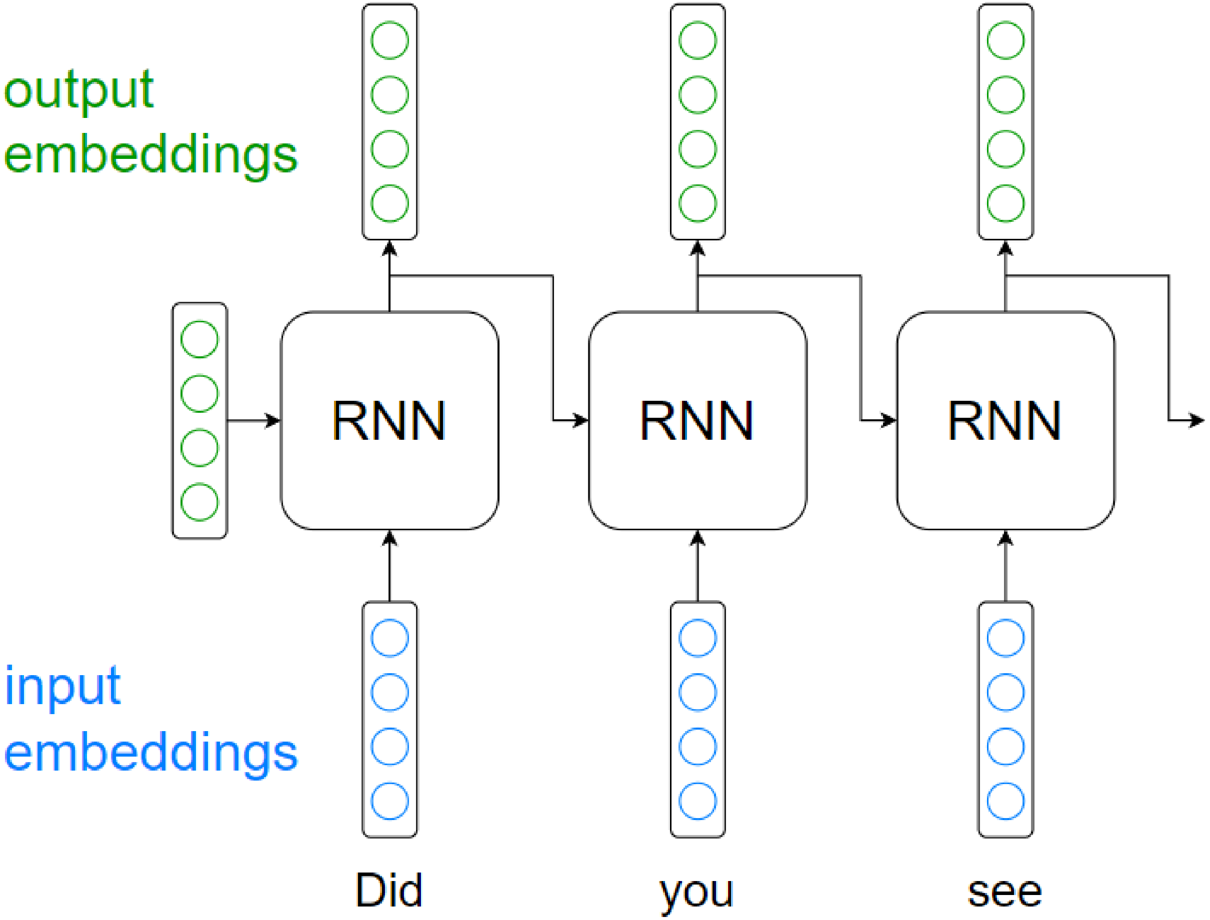
RNN for language modeling



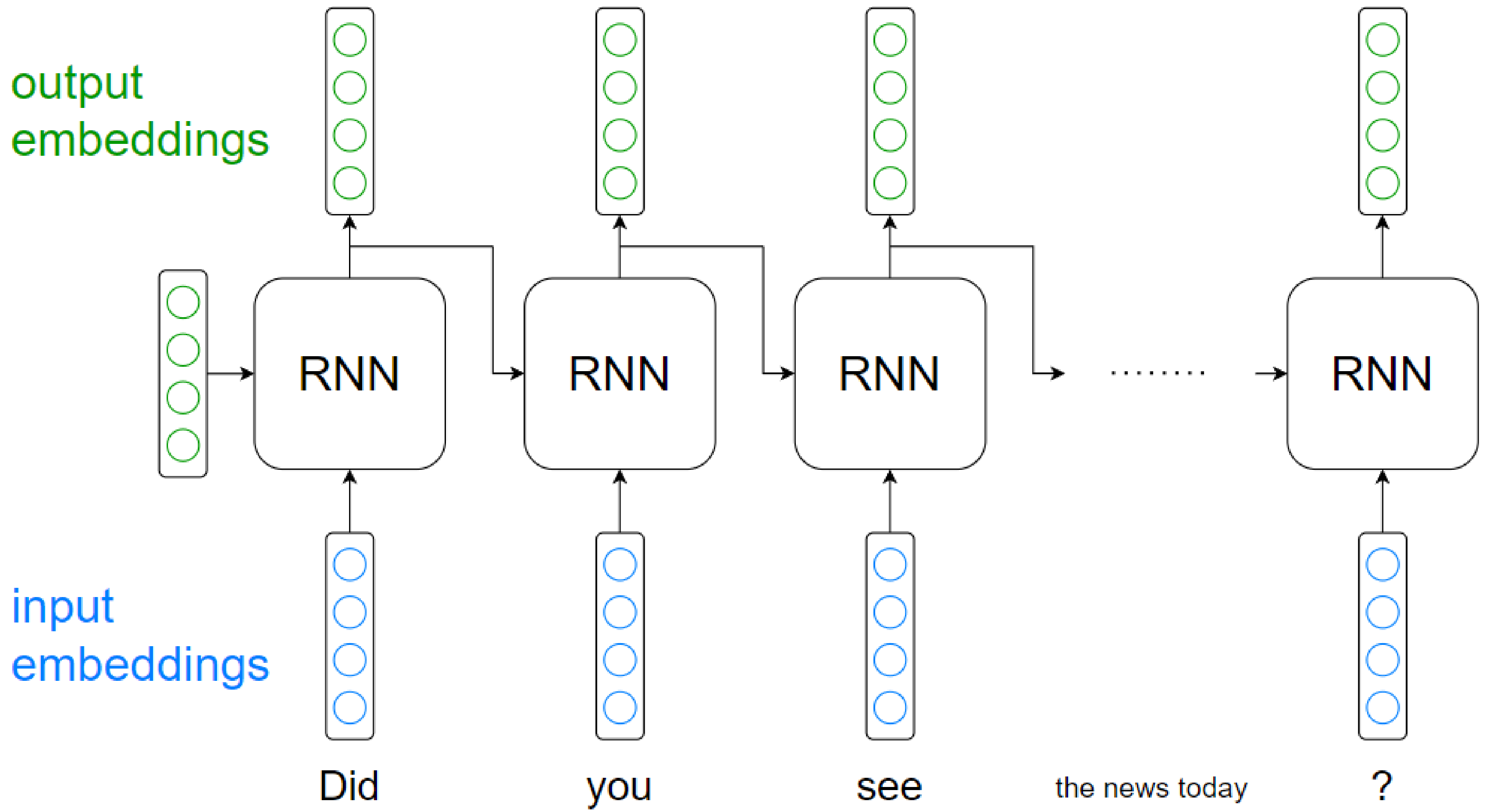
RNN for language modeling



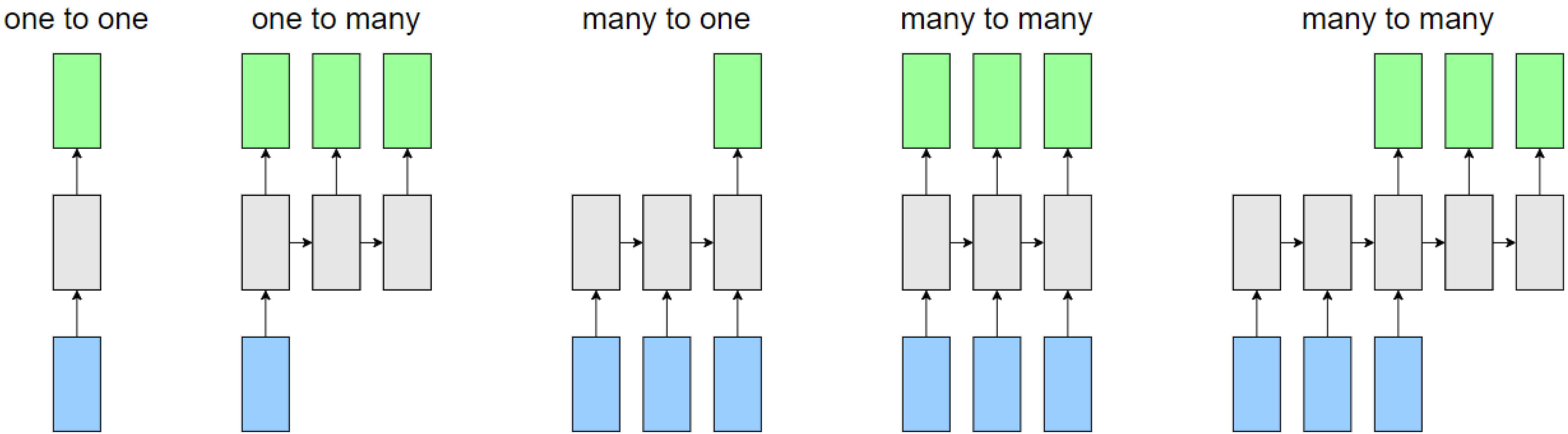
RNN for language modeling



RNN for language modeling

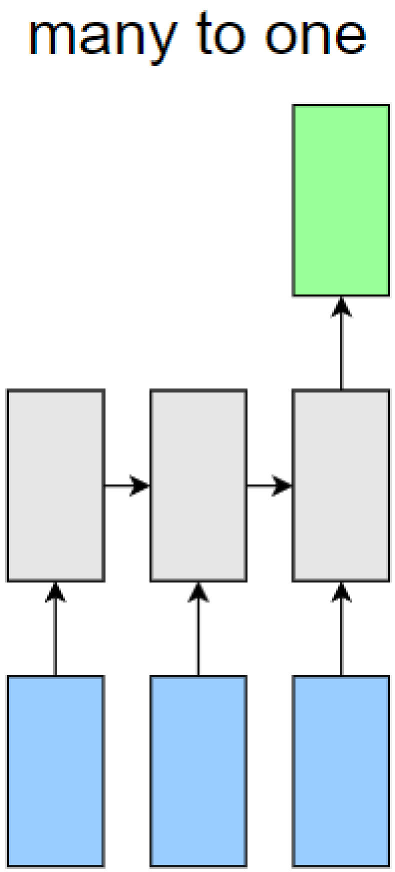


RNN – processing sequences



RNN – many to one

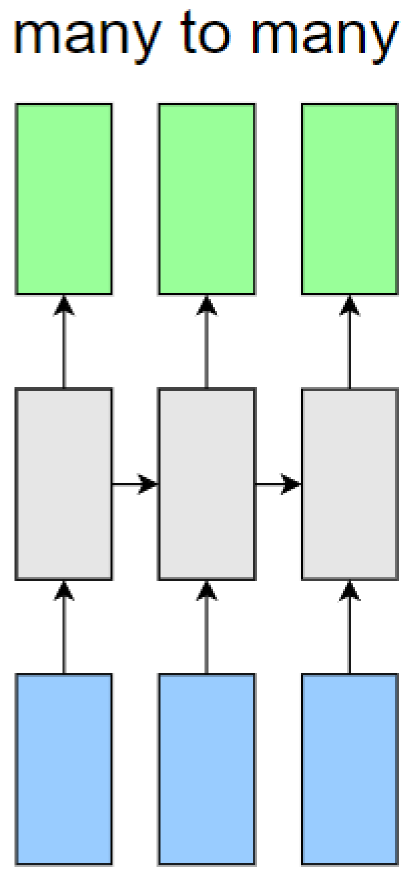
- Text classification
  - News article → topic
  - Customer review → positive / negative sentiment
  - Email → spam / legit



RNN – many to many

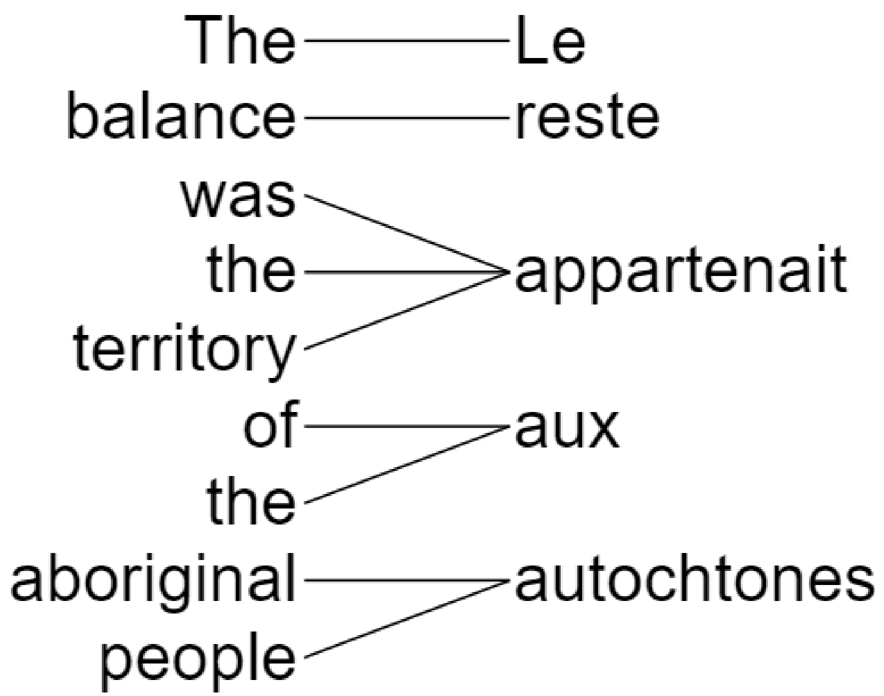
- Annotation tasks
    - Spell checking
    - Named entity recognition (NER)
    - POS checking
    - etc.

- Translation (e.g. English to French)?



Machine translation

- Alignment is not 1:1
- Word order may change



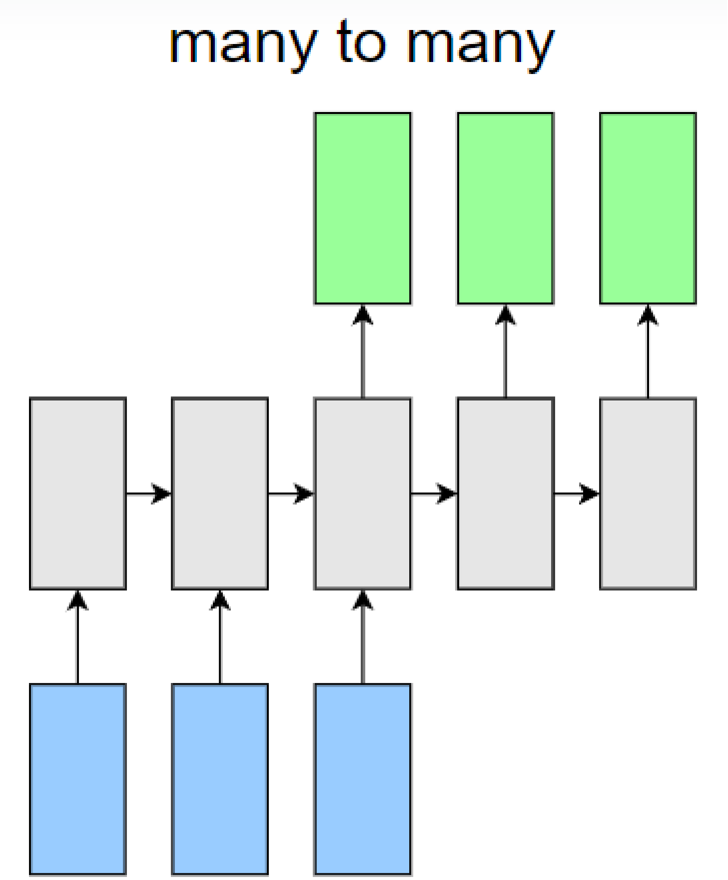
Brown, Peter F., et al. "The mathematics of statistical machine translation: Parameter estimation." Computational linguistics 19.2 (1993): 263-311.

RISC Software GmbH

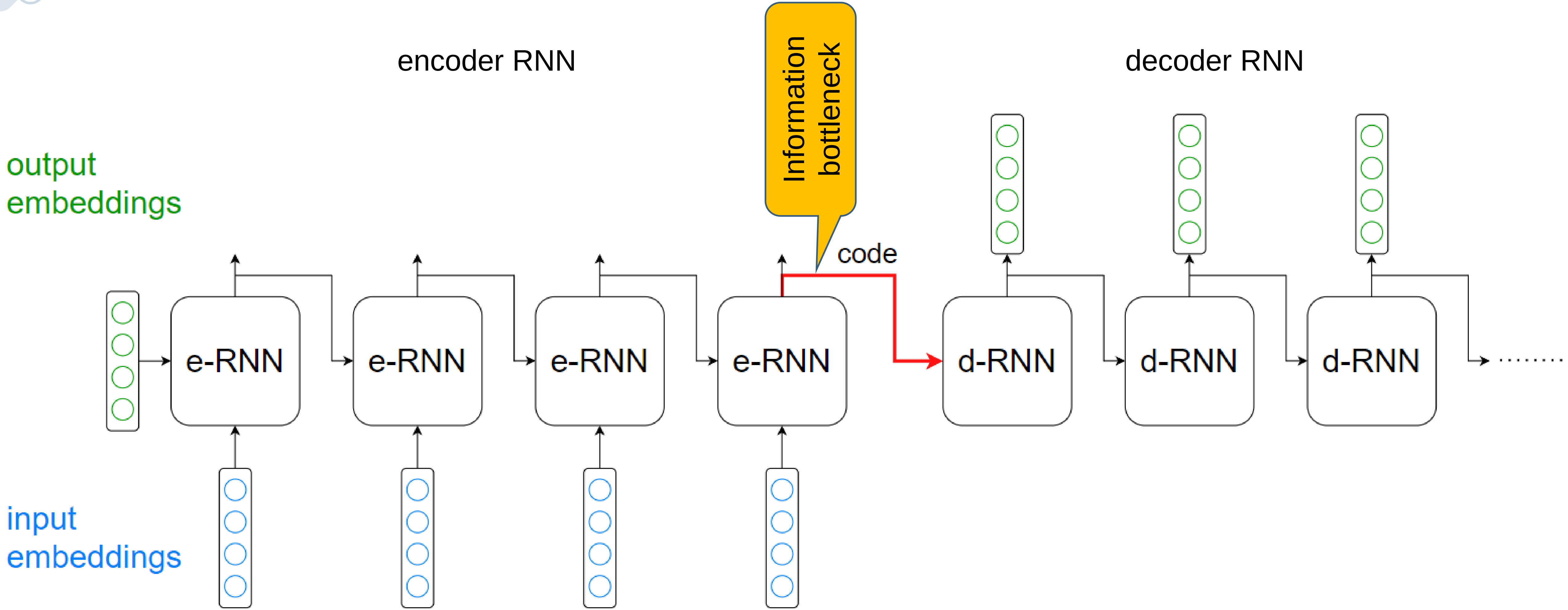
RNN – many to many

- Machine translation
  e.g. English → French

- Information extraction
  e.g. text summarization



RNN – many to many

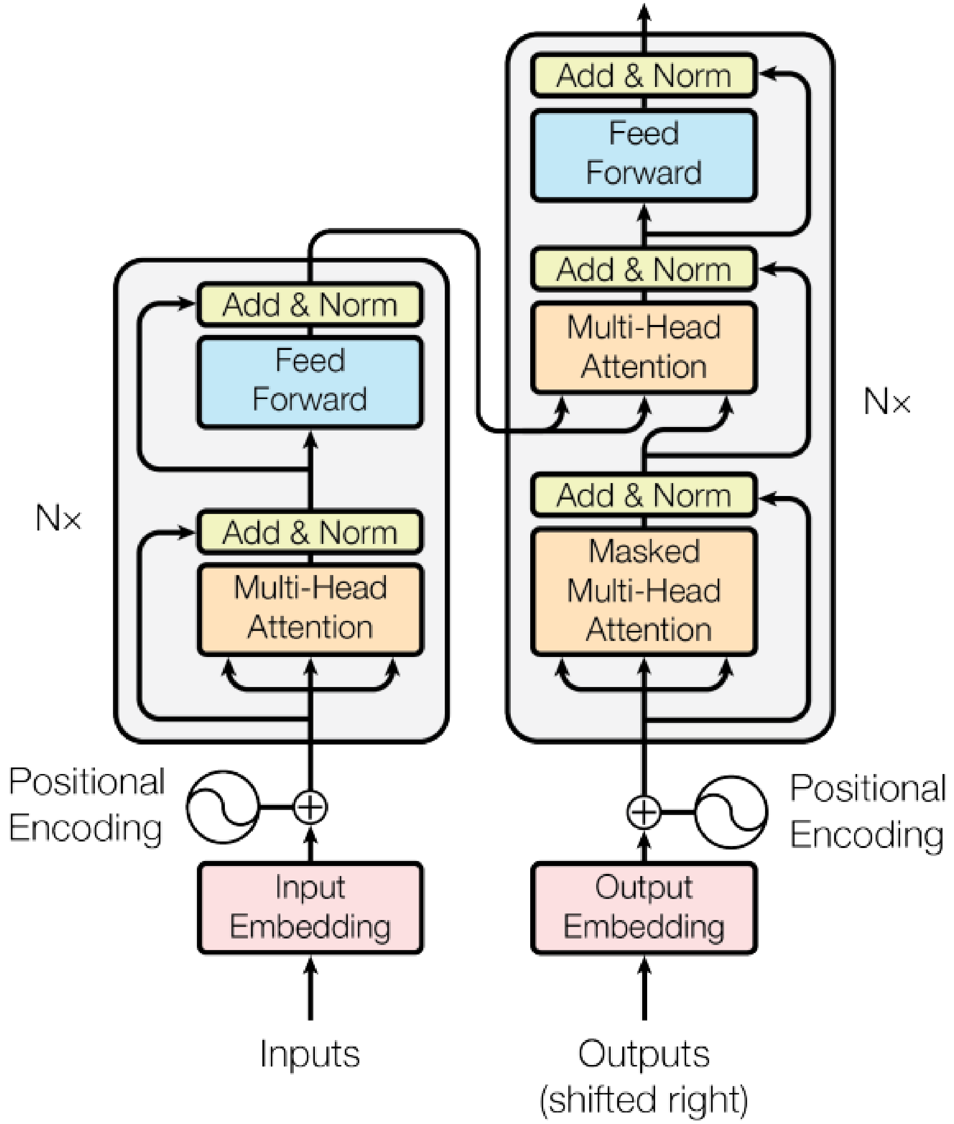


Transformers

- Transformer architecture (2017)
  - Solve information bottleneck
  - Multi-head attention
  - no RNN
  - DL best practices

- SotA for many NLP tasks

- Many implementations
  - BERT
  - T5
  - GPT-3



https://arxiv.org/pdf/1706.03762.pdf

Kontakt

RISC Software GmbH
Softwarepark 32a • 4232 Hagenberg • Austria
+43 7236 93028 • https://www.risc-software.at